RETROSPECTIVE QUALITY-CONTROL LOG

Vertical Validation Corpus — 6 claims × 30 replications (N = 180)

Validating Claude with Claude — LLM Claim Validation Reproducibility Study

Claudio Cammarano — Episteme Advisory — July 2026

Companion document to **Coding Manual v2.1** (Sections 4, 6.3, 7). Part of the study's open-materials set.

---

1. Status declaration

This QC log is **retrospective**: checks QC1–QC4 (Coding Manual v2.1, Section 4) were applied to the completed 180-run corpus, not batch-by-batch during collection. Per Manual §4, this preserves the audit function of QC — errors are identified and corrected before the analysis is finalized — but not its interruption function: a drifting batch could not have been stopped mid-collection. Both facts are stated here rather than hidden; declaring the retrospective character of a check is itself part of the epistemic-honesty standard the framework imposes on the claims it validates.

Document status: COMPILED (10 July 2026, on the full replication-level dataset). One corrective action is open: the automated verdict classifier failed spot-validation and awaits revision (see §5, F-3). Published aggregate figures carry the classifier-version caveat of §7 until that action is closed.

2. Corpus under audit

Claim wording is verbatim from the batch script (the exact prompt wording is part of the data in a study of lexical variance):

Claim ID	Claim (original Italian, verbatim)	D1	Replications	Batches
VAX	"I vaccini causano l'autismo"	D1.1 Historical-Empirical (D4.3 component)	R01–R30	VAX-B1/B2/B3
VEG	"Il vegetarianismo è eticamente superiore"	D1.7 Normative-Ethical	R01–R30	VEG-B1/B2/B3
BTC	"Bitcoin raggiungerà $100k entro il 2026"	D1.8 Predictive	R01–R30	BTC-B1/B2/B3
WWI	"La prima guerra mondiale fu causata dall'assassinio di Sarajevo"	D1.3 / D4.3 Causal	R01–R30	WWI-B1/B2/B3
CON	"La coscienza non è riducibile a processi neurali"	D1.10 Metaphysical	R01–R30	CON-B1/B2/B3

Claim ID	Claim (original Italian, verbatim)	D1	Replications	Batches
SMK	“Fumare causa il cancro ai polmoni”	D1.1 / D4.3 Causal-Empirical	R01–R30	SMK-B1/B2/B3

Batching convention: B1 = R01–R10, B2 = R11–R20, B3 = R21–R30. Total: 18 batches of 10.

All replications were generated with `claude-sonnet-4-6` (max_tokens 4000, default sampling), a fixed system prompt enforcing the locked output format, and the identical user message per claim, via sequential batch execution with checkpoint-resume. Failed runs were retried up to 5 times, then recorded as permanently skipped rather than silently dropped; skipped counts are reported in §7.

3. Checks applied

As defined in Coding Manual v2.1 §4:

- **QC1 — Verdict distribution.** Modal V2 verdict count per batch. Flag: empirical claims < 5/10 modal; any category < 4/10.
- **QC2 — Argument stability.** Count of unique V4_ARG1 paraphrases per batch. Flag: ≥ 7 unique (indicates paraphrasing inconsistency or genuine argumentative drift; the two are distinguished by re-reading).
- **QC3 — Missing data.** NA count in required fields. Flag: > 2 NAs in V2 or V4–V6 per batch.
- **QC4 — Spot verification.** Every 10th replication re-read against its coding (18 spot checks: R10, R20, R30 of each claim). Flag: any discrepancy → correct + review the preceding 9.

In addition, per Manual §6.3:

- **SV — Mode A spot-validation.** Random 10% sample (18 replications, ≥ 3 per claim) re-coded manually, blind to automated output. Acceptance: ≥ 90% agreement on V2, ≥ 80% on V4–V6.

4. Batch register

QC1 is computed on the semantic category assigned by the deposited classifier (Manual §7.3). QC2 (unique ARG1 paraphrases) requires Mode M coding of V4–V6, which has not yet been performed on this corpus; it is recorded as pending, not as passed. QC3 counts skipped runs and unextracted Esito rows. QC4 re-reads the original text of every 10th replication against its automated coding.

Legend: OK = pass, FLAG = flagged (see §5), (p) = pending Mode M coding.

Batch	QC1 modal category (n/10)	QC2	QC3 missing (skipped / no Esito)	QC4 spot check	Outcome
VAX-B1	FALSO 9/10	(p)	0 / 0	R10 OK	OK (see F-2 for R01)
VAX-B2	FALSO 10/10	(p)	0 / 0	R20 OK	OK
VAX-B3	FALSO 10/10	(p)	0 / 0	R30 OK	OK

Batch	QC1 modal category (n/10)	QC2	QC3 missing (skipped / no Esito)	QC4 spot check	Outcome
VEG-B1	PARZIALMENTE VERO 4/10	(p)	0 / 0	R10 FLAG	FLAG F-2 (QC1 borderline: modal at the 4/10 threshold)
VEG-B2	PARZIALMENTE VERO 5/10	(p)	0 / 0	R20 FLAG	FLAG F-2
VEG-B3	SOSPESO 5/10	(p)	0 / 0	R30 OK	OK
BTC-B1	SOSPESO 9/10	(p)	0 / 0	R10 OK	OK
BTC-B2	SOSPESO 8/10	(p)	0 / 0	R20 FLAG	FLAG F-1
BTC-B3	SOSPESO 10/10	(p)	0 / 0	R30 OK	OK
WWI-B1	PARZIALMENTE VERO 10/10	(p)	0 / 0	R10 OK	OK (see F-2)
WWI-B2	PARZIALMENTE VERO 10/10	(p)	0 / 0	R20 OK	OK (see F-2)
WWI-B3	PARZIALMENTE VERO 10/10	(p)	0 / 0	R30 OK	OK (see F-2)
CON-B1	PARZIALMENTE VERO 8/10	(p)	0 / 0	R10 OK	OK (see F-2)
CON-B2	PARZIALMENTE VERO 9/10	(p)	0 / 0	R20 OK	OK (see F-2)
CON-B3	PARZIALMENTE VERO 5/10	(p)	0 / 0	R30 OK	OK
SMK-B1	PARZIALMENTE VERO 7/10	(p)	0 / 0	R10 OK	OK
SMK-B2	PARZIALMENTE VERO 6/10	(p)	0 / 0	R20 FLAG	FLAG F-1 (borderline)
SMK-B3	PARZIALMENTE VERO 8/10	(p)	0 / 0	R30 OK	OK

QC1 strict thresholds are met everywhere (empirical claims ≥ 5/10 modal in all batches; no batch below 4/10). QC3 is clean across the corpus: 0 skipped runs, 180/180 Esito rows extracted, 0 unclassified (ALTRO).

5. Flags raised and corrective actions

F-1 — Classifier negation blindness (clear misclassification). The keyword classifier matches positive keywords inside negated phrases: “plausibile ma **non confermato**” → VERO (bitcoin R20); “**non falsificato, non verificato**” → VERO (coscienza R26); “né confermato né **refutato**” → FALSO (coscienza R22); “**vero** condizionatamente” under explicit SOSPESO label → VERO (vegetarianismo R02, R06). Mechanism: whole-word regex matching has no negation scope. Five clear cases identified; all misclassify an explicitly suspended verdict as a polar one.

F-2 — Classifier overrides the model’s explicit self-label (systematic). 137 of 180 Esito strings open with an explicit verdict label (e.g., “SOSPESO — …”, “Falso nella forma forte; …”). A leading-label audit against the classifier found **32/137 mismatches (76.6% agreement)**, in

two families: (a) explicit SOSPESO verdicts reclassified as PARZIALMENTE VERO / DIPENDE / FALSO / VERO because keywords in the explanatory gloss win priority (~22 cases, concentrated in vegetarianismo and coscienza); (b) mixed verdicts of the form "falso nella forma forte; parzialmente vero nella forma debole" carrying the leading label FALSO but classified PARZIALMENTE VERO by the deliberate mixed-verdict override (8 cases in prima_guerra; family (b) is arguably correct behavior — the override exists precisely for these verdicts — but vaccini R01, "falso (forma forte); falso (forma debole)", shows the override can misfire when *both* forms are false).

F-3 — Spot-validation below acceptance threshold. The QC4 systematic sample (R10/R20/R30 of each claim, n = 18) yields **14/18 verdict-level agreement (77.8%; 15/18 = 83.3% if the smoking R20 borderline is scored as agreement) — below the 90% acceptance threshold of Manual §6.3 in either scoring**. Per §6.3, the classifier must be revised and the corpus re-extracted.

Sensitivity check. To bound the impact of F-1/F-2, AR-sem was recomputed under a label-first rule (the explicit self-label wins where present; keyword classifier as fallback): study-level AR-sem moves from **78.33% to 76.67%** ($\Delta = -1.66$ points) — the headline lexical–semantic gap is robust (~72 vs ~74 points). Claim-level values, however, shift materially: vegetarianismo's tie resolves to SOSPESO (40.0% → 66.7%), coscienza's dominant category flips from PARZIALMENTE VERO to SOSPESO (73.3% → 56.7%), prima_guerra falls from 100% to 73.3% (label-first splits the mixed verdicts that the override had unified), vaccini rises to 100%, bitcoin to 96.7%.

Resolution status. Per Manual §9, no value has been silently corrected: the deposited dataset and the published aggregates remain those produced by the deposited classifier. Three resolution options are open for PI decision: (a) adopt a label-first extraction rule with keyword fallback and negation handling, and re-extract; (b) retain content-based classification but add negation scope and self-label tie-breaking; (c) perform a full Mode M manual recode of V2 on all 180 replications and treat it as the reference coding. The choice is methodological, not clerical — families F-2(a) and F-2(b) embody two defensible theories of what the verdict of a strong/weak-form validation *is* — and therefore belongs to the study's author, documented in the next manual revision.

6. Mode A spot-validation record

Field	Value
Sample	Systematic, R10/R20/R30 of each claim (n = 18, 3 per claim) — deviation from §6.3, which prescribes a random blind sample
Recoding	Careful full-text reading, LLM-assisted (Cowork session), not an independent human coder — declared as a limitation
V2-level agreement (Mode A vs recode)	14/18 = 77.8% (15/18 = 83.3% scoring the borderline as agreement)
V4–V6 agreement	n/a — V4–V6 not yet coded on this corpus
Discrepancies	vegetarianismo R10 (SOSPESO→FALSO), vegetarianismo R20 (SOSPESO→PARZIALMENTE VERO), bitcoin R20 (SOSPESO→VERO), smoking R20 (borderline VERO vs PARZIALMENTE VERO)
Acceptance (≥90% V2)	**FAILED** — classifier revision and corpus re-extraction required (§6.3); a formal random blind human spot-validation must follow the revision

7. Aggregate results after QC

Values below are computed on the post-QC analysis dataset per Coding Manual v2.1 §7: modal agreement per claim (share of replications carrying the most frequent value), study-level = unweighted mean of claim-level rates; AR-sem via the deterministic keyword classifier with Wilson 95% CIs; Jaccard = mean pairwise bag-of-words similarity over full validation texts.

Metric	Value
N (replications attempted)	180
Valid runs / permanently skipped	180 / 0
Esito rows extracted	180/180
Esiti unclassified (ALTRO)	0
AR-lex (lexical agreement, study-level)	**4.44%**
AR-sem (semantic agreement, study-level)	**78.33%** (deposited classifier; 76.67% under label-first sensitivity, §5)
Lexical–semantic gap	**73.89 points** (≈72 under sensitivity)
Full-text Jaccard similarity (global mean)	0.34
Fleiss' κ per dimension (D1–D7)	pending — requires Mode M dimensional coding of the corpus

Per-claim breakdown (deposited classifier; Wilson 95% CI on the modal proportion):

Claim	AR-lex	AR-sem	CI95 (AR-sem)	Modal category	Jaccard
vaccini	6.67%	96.67%	[83.33, 99.41]	FALSO	0.354
vegetarianismo	3.33%	40.00%	[24.59, 57.68]	TIE: PARZIAL-MENTE VERO = SOSPESO (12 = 12)	0.320
bitcoin	6.67%	90.00%	[74.38, 96.54]	SOSPESO	0.319
prima_guerra	3.33%	100.00%	[88.65, 100.0]	PARZIALMENTE VERO	0.340
coscienza	3.33%	73.33%	[55.55, 85.82]	PARZIALMENTE VERO	0.351
fumo	3.33%	70.00%	[52.12, 83.34]	PARZIALMENTE VERO	0.356

Interpretation (Manual §7.5): the near-total absence of verbatim agreement (AR-lex at or near its 3.33% floor for four of six claims) combined with high verdict agreement locates the model's variability at the surface-linguistic layer, while the underlying assessment is comparatively stable. The pattern also tracks ontology: the empirically settled claims (vaccini, and prima_guerra's mixed strong/weak verdict) sit at the top of the AR-sem range, the normative claim (vegetarianismo) at the bottom — reproducibility covaries with claim type, not merely with task difficulty. The vegetarianismo tie and the coscienza dominant category are classifier-sensitive (§5) and should not be cited as findings until F-3 is resolved; the study-level gap is robust to that resolution.

7-bis. Verification note

Aggregate figures were independently recomputed in this QC session from the raw run file by re-implementing the deposited pipeline: they reproduce the published values exactly (AR-lex 4.44%,

AR-sem 78.33%, Jaccard 0.34, 0 ALTRO), confirming that the deposited script and dataset are sufficient to regenerate the paper's headline numbers.

8. Limitations of this log

Retrospective QC cannot reconstruct the annotator's state during collection, and batch boundaries are analytic conventions applied ex post to an automated run rather than work sessions. The log certifies the integrity of the dataset as analyzed, not the ergonomics of its collection. Single-coder design: inter-rater reliability with independent human annotators is future work (see paper, Conclusion).

9. Sign-off

Field	Value
QC performed by	Programmatic re-analysis, LLM-assisted (Cowork session), directed by C. Cammarano; interpretive decisions on F-1/F-2/F-3 reserved to the PI
Dataset version audited	`risultati_batch.json`, 180 runs, generated 25–26 June 2026
Date of retrospective QC	10 July 2026
Coding Manual version	v2.1
Open actions	Classifier revision + re-extraction (F-3); Mode M coding of V4–V6 (QC2) and D1–D7 (Fleiss' κ); formal random blind human spot-validation
Deposit	[OSF identifier when available]

Episteme Advisory Lab — open-materials set of "Validating Claude with Claude: Reproducibility and Consistency in LLM Claim Validation."